http://dx.doi.org/10.7236/JIIBC.2013.13.4.141
JIIBC 2013-4-19

특징선택과 특징가중의 융합을 통한 웹문서분류 성능의 개선

Performance Improvement of Web Document Classification through Incorporation of Feature Selection and Weighting

이아람*, 김한준**, 현만***

Ah-Ram Lee, Han-Joon Kim, Xuan Man

요 약 기계학습을 이용한 자동분류시스템은 학습과정을 통해 분류모델을 구축하고 이를 기반으로 미분류 데이터를 특정 카테고리로 분류한다. 기계학습 기반 자동분류 시스템의 성능은 분류모델의 구성 인자인 특징의 품질에 크게 의존한다. 문서 데이터의 경우 특징 집합을 생성하기 위해 문서내의 출현단어와 문서의 구조적 정보를 활용한다. 특히 웹문서로부터 특징을 추출하기 위해 단어뿐만 아니라 태그, 하이퍼링크 정보를 분석할 수 있다. 최근 웹문서의 분류 기법에 대한 연구는 기계학습 알고리즘보다 특징 생성 및 가공 기술에 초점을 맞추고 있다. 이에 본 논문은 웹문서의 분류모델을 개선하기 위해 단어, 태그, 하이퍼링크 정보로부터 고품질의 특징을 선별 추출하여 가중치를 자동으로 부여하는 기법을 제안한다. Web-KB 문서집합을 이용한 다양한 실험을 통해 제안 기법의 우수성을 보인다.

Abstract Automated classification systems which utilize machine learning develops classification models through learning process, and then classify unknown data into predefined set of categories according to the model. The performance of machine learning-based classification systems relies greatly upon the quality of features composing classification models. For textual data, we can use their word terms and structure information in order to generate the set of features. Particularly, in order to extract feature from Web documents, we need to analyze tag and hyperlink information. Recent studies on Web document classification focus on feature engineering technology other than machine learning algorithms themselves. Thus this paper proposes a novel method of incorporating feature selection and weighting which can improves classification models effectively. Through extensive experiments using Web-KB document collections, the proposed method outperforms conventional ones.

Key Words : Document Classification, Web, Feature Selection, Feature Weighting, Machine Learning

Ⅰ. 서 론

최근 IDC 디지털 유니버스 연구보고서에 따르면, 2011년 생성된 데이터의 양은 약 1.8 Zettabytes (1.8조 Gigabytes)로 추정하며, 향후 10년 동안 그 규모는 50배를 초과할 것이며, 그 중에서 비정형(unstructured) 또는

*정회원, 서울시립대학교 전자전기컴퓨터공학부
**정회원, 서울시립대학교 전자전기컴퓨터공학부 (교신저자)
***준회원, 서울시립대학교 전자전기컴퓨터공학부
접수일자 : 2013년 7월 9일, 수정완료 : 2013년 8월 9일
게재확정일자 : 2013년 8월 16일

Received: 9 July 2013 / Revised: 9 August, 2013
Accepted: 16 August, 2013
**Corresponding Author: khj@uos.ac.kr
School of Electrical and Computer Engineering, University of Seoul, Korea

반정형(semi-structured) 데이터가 90%에 달할 것이라는 전망이다[1]. 의미있는 대다수의 정보는 비·반정형적 텍스트의 형태로 존재하기 때문에 빅데이터(big data) 시대에 텍스트의 자동분류(text classification) 기술은 매우 중요하게 다뤄지고 있다. 자동분류는 주어진 문서에 대하여 기 정의된 카테고리(category)에 자동으로 분류하는 기술을 말한다. 특히 모바일 기기의 확산에 따라 소셜 미디어 내부에 잠재적 가치가 큰 텍스트 데이터가 엄청난 규모로 생산되고 있어 자동분류기술의 중요도는 날로 커지고 있다[2][3][4][5].

최근 대부분의 자동문서분류 기법은 기계학습(machine learning) 알고리즘을 활용한다. 이것의 기본 구조는 학습(learning) 단계와 분류(classification) 단계로 구성된다. 학습 단계에서 학습문서 집합으로부터 각 카테고리에 출현하는 특징(feature) 집합을 주요 인자로 하여 카테고리를 분별할 수 있는 '분류모델(classification model)'을 구축한다. 이를 사용하여 분류 단계에서 미분류 문서에 대한 분류 작업을 수행한다. 문서 데이터에 주로 적용되는 알고리즘은 나이브베이즈(Naïve Bayes), 서포트벡터머신(support vector machine), k-최근접이웃(k-nearest neighbors) 등이다. 특히 나이브베이즈 알고리즘은 분류모델의 단순함에 비해 성능이 우수한 것으로 평가되어 자동분류시스템에 자주 활용되고 있다[6]. 본 연구에서도 이를 채택하여 제안 기법의 우수성을 보인다.

일반적으로 텍스트 문서에 대한 자동분류시스템은 그 성능이 학습 알고리즘 자체보다는 특징선택(feature selection) 알고리즘에 의존하는 경향이 크다[7][8]. 특징선택이란 학습문서에 존재하는 특징(또는 단어)들이 지나치게 많아 이 중에서 카테고리 간 차별화에 기여하는 특징만을 골라내는 기법을 의미한다. 비근한 예로 문서집합에 포함된 모든 단어를 분류 모델의 생성에 사용한다면 분류 클래스의 특성을 반영하지 않은 단어가 다수 포함되어 분류 성능을 저해할 뿐만 아니라 학습 시간을 크게 연장시키는 결과를 초래한다. 자동문서분류 시스템을 구성하는데 있어서 특징선택을 통해 특징공간의 복잡도를 줄이고 분류에 기여하지 않는 특징들을 삭제하는 과정은 필수적이다.

더 나아가 본 연구는 하이퍼텍스트(hyptertext) 특성을 가지는 웹문서에 대한 자동분류의 성능 개선을 위해서 특징선택과 특징가중 기법에 초점을 맞춘다. 웹문서는 단어 이외에도 하이퍼링크(hyperlink)정보, 태그(tag) 정보 등의 구조적 정보를 갖는다. 특히 하이퍼링크는 웹문서 간의 관계를 나타내는 중요한 정보이며, 링크로 연결된 문서들은 서로 유사한 내용을 공유하고 있거나 연관되는 정보를 가지는 것이 보통이다. 이러한 특성을 반영하여 링크로 연결된 웹문서의 단어를 특징으로 이용하거나, 웹문서 간의 관계를 고려하여 특징에 가중치를 부여하여 고품질의 특징을 선택함으로써 문서분류의 성능을 향상시킬 수 있다.

본 논문의 구성은 다음과 같다. 우선 2절에서 웹문서에 포함된 태그 및 하이퍼링크 정보에 관한 배경 지식 및 이를 이용한 관련 연구를 살펴보고, 3절에서는 그러한 구조적 정보를 이용하여 선별된 특징에 대하여 최적의 가중치를 부여하는 방법을 제안한다. 4절에서는 제안한 기법에 대한 실험적 평가를 제시하고 5절에서는 결론과 향후 연구 방향을 제시한다.

II. 배경 지식 및 관련 연구

웹문서는 월드와이드웹(World Wide Web)상에 유포되는 문서로서 하이퍼텍스트 마크업 언어(HyperText Markup Language, 이하 HTML)를 사용하여 작성된다. HTML은 웹문서를 웹브라우저에 효과적으로 보여주기 위하여 제목, 단락, 목록, 테이블 등과 같은 구조적 형태와 특정 텍스트 영역에 대한 다양한 표현을 지원하기 위해 여러 유형의 태그를 포함한다. 예를 들어 웹문서의 제목은 <title>, 부제는 <h1>, 실제 내용은 <body>, 단락은 <p>, 목록은 <li>, 테이블은 <td>, 그리고 하이퍼링크 정보를 가진 단어, 즉 앵커 텍스트(anchor text)는 <a> 라는 태그를 사용한다. 특히 하이퍼링크 태그는 웹문서간 연결을 지원하는 중요한 요소이다. 이러한 태그 정보는 웹문서에 포함된 단어의 중요도를 평가할 수 있는 기준이 될 수 있다. 예를 들어, <title>, <h1>, <a> 등의 태그에 포함된 단어는 그렇지 않은 단어와 비교하여 상대적으로 높은 중요도를 부여할 수 있다.

우선 웹문서간의 하이퍼링크 관계 및 관련 용어를 살펴보자. 연결 관계의 기준이 되는 웹문서를 '기준 웹문서'라 하고, 기준 웹문서와 직접으로 링크하고 있는 웹문서를 '인접 웹문서'라 한다. 여기서 기준 웹문서를 링크하고 있는 웹문서를 '선임문서'라 하며, 이는 기준 웹문서와 '진입링크(incoming link)' 관계를 맺고 있는 것이다. 반

대로 기준 웹문서가 링크하고 있는 문서를 '후임문서'라 하며, 이는 기준 웹문서와 '진출링크(outgoing link)' 관계를 맺고 있는 것이다.

웹문서의 경우, 분류모델을 생성하기 위해 문서에 출현하는 단어뿐만 아니라 인접 웹문서가 가진 정보를 아울러 사용할 수 있다. 예를 들어, 인접 웹문서의 카테고리 정보를 이용하여 기준 문서의 분류 정확도를 높이려한 연구가 있다[9]. 이는 기준 문서의 카테고리를 결정하기 위해 k-최근접이웃 학습 알고리즘과 유사하게 인접문서의 카테고리 정보를 활용한다. 하지만 이는 기준 문서가 항상 인접 문서와 카테고리가 동일하지 않아 분류 성능이 좋지 않다. 인접 문서의 정보를 보다 적극적으로 활용하는 하나의 방안이 인접 문서에 포함된 단어를 마치 기준 문서에 포함된 단어인 것으로 간주하는 것이다[10]. 그런데 인접 문서의 모든 단어를 특징으로 이용한다면 불필요한 특징 정보를 과도하게 포함할 수 있어서 적절한 부분집합을 추려내는 것이 주요 관건이다.

추가적으로 본 연구에서는 인접 정보를 이용하는 또 하나의 방안으로서 기준 문서의 권위도(prestige)를 이용하고자 한다. 즉 권위도가 높은 웹문서에 포함된 단어는 권위도에 비례하여 가중치를 부여하는 것이다. 권위도는 하이퍼링크 구조를 가지는 웹문서와 같이 인용과 참조의 의미로 연결된 데이터 집합에서 상대적 중요도를 평가하는 정량적 척도가 된다[11]. 이와 관련한 대표적 알고리즘이 Google 검색엔진(www.google.com)의 기반이 된 페이지랭크(PageRank)[11]이다. 이는 상대적으로 중요한 웹문서는 많은 사이트로부터 진입링크를 받는다는 관찰 결과와 웹문서의 하이퍼링크를 따라 랜덤하게 탐색하여 권위도를 계산하는 랜덤워크(random walk) 모델에 바탕을 두고 있다. 결론적으로 권위도가 높은 웹문서일수록 중요도가 높은 문서가 된다. 그런데 페이지랭크 알고리즘은 랜덤워크 모델에 따라 거미줄처럼 얽혀 있는 링크 관계에서 상호간에 권위도(즉 페이지랭크) 값을 주고받으면서 수렴할 때까지 반복적으로 계산해야 하기 때문에 시간 복잡도가 매우 높다. 본 연구에서는 검색이 아닌 자동분류를 목적으로 분류모델을 구축하기 때문에 시간 복잡도를 줄이기 위하여 단순한 권위도 계산법을 제안한다.

Ⅲ. 인접 웹문서를 이용한 특징선택과 가중치 부여 기법

앞서 기술한 바와 같이 본 연구는 웹문서의 자동분류의 성능을 높이기 위해 태그 및 하이퍼링크라는 구조적 정보를 적극 활용한다. 본 절에서 태그 및 하이퍼링크 정보를 활용하여 특징선택의 문제와 선정된 특징에 대한 특징가중의 문제를 동시에 다루는 기법을 소개한다.

1. 태그 정보를 이용한 특징 추출

기계학습 기반 분류모델를 구성하는 주요 인자는 특정 카테고리에 속한 각 특징의 출현 빈도수가 되며, 결국 그것의 중요도는 출현 빈도수에 비례한다. 그런데 웹문서의 주요 태그(예를 들어 <title>)에 출현하는 단어는 그 출현 횟수가 적은 경우 정성적으로 그 중요도가 크다 할지라도 분류모델의 주요 특징이 되지 못할 수 있다. 일반적으로 <title> 태그에는 본문을 상징하거나 대표하는 단어들이 포함된다. 그러므로 주요 태그에 속한 단어는 다른 단어들보다 그 중요도가 높게 설정되는 것이 바람직하다. 즉 웹문서의 주요 태그에 속한 단어는 해당 문서의 중요도를 결정하기 때문에 그 태그에 포함되지 않은 단어보다 분류모델을 위한 주요 특징으로 사용되는 것이 바람직하다[12].

특히 하이퍼링크에 포함된 앵커텍스트(anchor text)는 기준 문서의 성격을 결정짓는 주요 단서가 된다. 앵커텍스트는 웹문서를 연결시켜주는 링크를 가진 단어로서 연결 관계를 정의하는 단어 집합이다. 이는 기준 웹문서의 내용을 담고 있으면서 인접 문서의 내용도 포함하는 단어들이기 때문에 문서분류를 위한 특징으로서 유용하게 활용될 수 있다. 인접 문서의 단어를 특징으로 이용한 여러 연구 중에서 앵커텍스트를 포함하여 전후 10개의 주변 단어들을 특징으로 사용한 기법이 가장 좋은 분류 성능을 보였다[10]. 이에 본 연구에서도 진입문서의 앵커텍스트 단어 및 이와 관련성이 높을 것으로 예상되는 주변 단어를 기준 문서의 특징에 포함시킨다.

요약하면 본 연구에서는 기준 문서의 특징집합에 해당 문서에 포함된 단어뿐만 아니라 그것의 인접문서에 존재하는 제목 태그 <title>의 단어, 하이퍼링크 태그 <a>의 단어, 그리고 기준 문서에 대한 진입문서의 하이퍼링크 태그 <a>의 주변 단어를 특징으로 사용한다.

2. 권위도를 이용한 특징가중

또한 하이퍼링크로 복잡하게 연결된 웹문서를 권위도에 따라 차등적으로 중요도를 계산하여 그 값을 특징가중 과정에 반영하여 분류모델의 성능 높이고자 한다. 기준 문서에 포함된 특징 단어의 가중치를 웹문서의 권위도에 비례하여 결정하자는 것이다. 권위도의 계산은 중요도가 높은 문서일수록 진입링크를 많이 받는다는 사실을 반영해야 한다. 예를 들어 페이지랭크 알고리즘에서는 문서 A가 문서 B, C, D에 대하여 총 3개의 하이퍼링크가 연결되었다면 문서 B는 문서 A가 가진 페이지랭크 값의 1/3 만큼을 가져온다. 이러한 원칙을 가지고 랜덤워크 모델에 따라 하이퍼링크 관계에 놓인 문서간에 페이지랭크 권위도 값을 주고받으면서 수렴할 때까지 반복적으로 계산한다. 결과적으로 기준문서가 진입링크를 받는 경우는 권위도가 증가하고, 반대로 다른 문서에 대하여 진출링크를 가지는 경우 권위도가 감소한다.

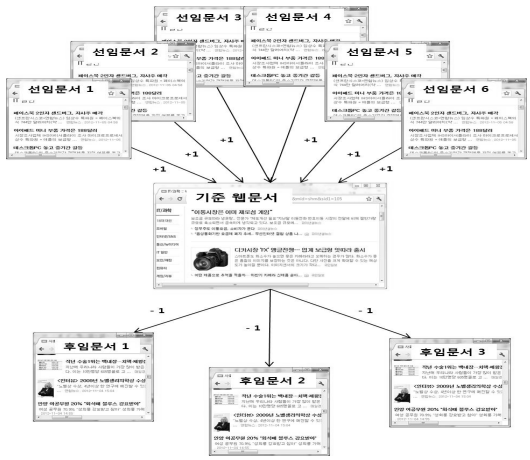

그림 1. 권위도 계산'의 예시
Fig. 1. Example of computing the prestige

본 연구에서는 권위도 계산의 목적이 분류모델을 구성하기 위한 특징가중(feature weighting)이기 때문에 대략적인 권위도를 추정하는 것으로 충분하다. 따라서 기준 문서를 중심으로 진입링크 개수와 진출링크 개수의 차이를 계산하여 그 값을 권위도로 추정에 이용하고자 한다. 기본적으로 기준 문서의 1개 진입링크에 대한 권위도 값을 +1로 설정한다. 이를 기준으로 1개 진출링크의 권위도 값 α는 -1에서 0까지의 범위로 설정한다. 예를 들어, 그림 1의 기준 문서는 6개의 진입링크와 3개의 진출링크를 가지고 있다. 진출링크의 중요도 α를 -1로 가정하여, 진입링크의 권위도는 +6, 진출링크의 권위도는 -3으로 평가하여 결과적으로 기준 문서의 권위도 값은 6+(-3) = 3이 된다. 이러한 계산법에 따라 페이지랭크와 유사하게, 다른 문서로부터 링크를 많이 받는 문서일수록 높은 권위도를 산출하게 된다. 해당 문서에 포함된 특징 단어는 권위도 값에 비례하여 가중치를 부여하게 되는 것이다. 상대적으로 진출링크가 많은 문서는 권위도 값이 음수가 될 수 있다. 0 이하의 권위도 값을 가진 문서는 중요도가 매우 낮은 것으로 평가하여 해당 문서에 포함된 단어는 기준 문서의 특징으로 사용하지 않는다.

3. 권위도 및 태그 정보를 융합한 특징가중 기법

특징가중 과정은 초기 주어진 특징 단어의 빈도수를 확장하는 것이다. 가중 결과로서 어떤 특징 단어의 빈도수가 과도하게 높아지면, 다른 중요한 특징의 빈도수를 상대적으로 낮추게 되어 오히려 분류모델의 성능을 저하시킬 수 있다. 그러므로 특징가중 과정은 보수적인 접근으로서 가중치를 제한된 범위 내의 값으로 변환시키는 것이 바람직하다.

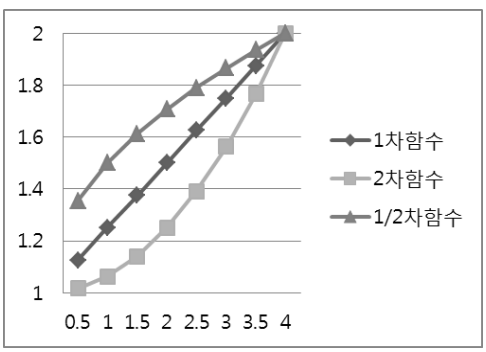

그림 2. 권위도의 변환
Fig. 2. Scaling of prestige values

초기 가중치의 산출을 위해 2절에 소개한 권위도 계산법에 따라 각 문서에 대한 권위도를 도출한다. 이때 권위도가 음수인 문서는 선택하지 않으므로, 권위도 P는 0이상의 값을 가진다. 앞서 언급한 바와 같이 이 수치를 바로 특징가중에 반영하기에는 변동성이 크므로, 정규화 차원에서 1부터 최대 M까지의 값으로 변환하여 이를 특징가중치 W_P로 취한다. 실험적으로 얻은 최적의 M값

은 2이다. 제안 기법에서는 그림 2에서 보는 바와 같이 권위도 P가 3가지 함수(1차 함수, 2차 함수, $\frac{1}{2}$차 함수)를 이용하여 권위도 기반 가중치 W_P가 계산된다. 기본적으로 1차 함수를 기준으로 $\frac{1}{2}$차 함수는 상대적으로 높은 가중치, 2차 함수는 상대적으로 낮은 가중치를 부여한다. 구체적으로 1차함수를 적용한 경우는, 권위도와 동일한 비율로 가중치를 높여준다. 2차 함수를 적용한 경우는, 권위도가 작을 때 가중치 W_P의 증가 비율을 줄이고 권위도가 커지면서 가중치 W_P의 증가 비율이 높아진다. 다시 말해서, 권위도가 높은 웹문서들의 가중치는 상대적인 차이를 높이고, 권위도가 낮은 웹문서들의 가중치는 상대적인 차이를 낮춘다. $\frac{1}{2}$차함수를 적용한 경우는 2차 함수를 적용한 경우와 반대의 성격을 가진다. 즉 권위도가 낮은 웹문서들의 상대적 가중치 차이가 권위도가 높은 웹문서들의 상대적 가중치 차이보다 커진다.

태그 정보에 기반한 가중 방안은 간단히 태그의 중요도에 따라 가중치 W_T를 결정한다. 본 논문에서는 제목 <title> 태그에 포함된 특징 단어에는 3, 앵커 태그 <a>에 포함된 특징 단어에는 2, 앵커텍스트 주변 특징 단어에는 1.5로 설정한다. 이는 태그의 의미적 중요도를 바탕으로 다수의 실험을 통하여 대략적으로 얻어낸 것이다. 보다 정밀한 가중치를 얻기 위해서는 담금질 기법(simulated annealing)과 같은 최적화(optimization) 기법을 적용해야 한다. 최적화 알고리즘은 다차원 문제의 공간에서 효율적으로 최적의 해를 탐색하는 기법을 말한다. 이를 분류모델의 성능의 개선되는 방향으로 각 태그에 대한 최적의 비중값을 추적해 나갈 수 있다. 최적화를 통한 가중치 결정 문제는 본 논문의 범위를 넘어서므로 이는 향후 연구에 포함시킨다.

결과적으로 특징가중에 사용되는 최종 가중치는 권위도 기반 가중치 W_P와 태그 기반 가중치 W_T를 융합한 값이 된다. 본 연구에서는 두 가중치를 단순히 합산하여 1에서 N까지 범위로 한정한다. 그리고 단어 빈도수는 양의 정수이어야 하므로 합산값을 반올림 처리한다. 그래서 최종 가중치 W는 식(1)과 같이 정의한다.

$$W = \left\lceil \frac{N-1}{4}(W_P + W_T) + \frac{5-N}{4} \right\rceil \quad (1)$$

여기서 N값은 $W_P + W_T$의 최대값인 5보다 작은 값을 취하게 되며, 실험적으로 2 근처의 값을 취할 때 분류모델의 정확성이 가장 높았다. 선정된 문서에 포함된 초기 특징 단어의 빈도수는 최종 가중치 W의 배수만큼 높여주게 된다. 예를 들어, 빈도수가 2인 특징 단어 t가 있을 때 이것의 최종 가중치 W가 2.5라면, 가중치를 적용한 특징 단어 t의 빈도수는 2×2.5 = 5가 된다. 특징 t의 빈도수는 5로 조정되어 분류모델의 구축에기여하게 된다. 위와 같이 특징선택과 특징가중의 융합 효과로 인해 권위도가 높은 문서에 포함된 특징 단어가 중요 태그에 출현한다면 그렇지 않은 특징 단어에 비해 상대적으로 높은 빈도수로 확장되어 결과적으로 분류모델을 개선하게 된다.

Ⅳ. 실험 및 결과

1. 실험 환경

제안 기법의 성능을 평가하기 위해서 본 연구는 Web-KB 웹문서집합을 이용하였다. 이는 4개 미국 대학교의 웹문서 8,282건를 수집한 것이며, 'student', 'faculty', 'staff', 'department', 'course', 'project', 'other' 등의 7개 카테고리로 구성되어 있다. 그리고 분류모델을 생성하기 위한 학습 알고리즘으로서 나이브베이즈 알고리즘을 채택하였다. 평가 척도는 분류정확도(classification accuracy)를 사용하며, 이는 분류한 문서 개수에 대한 옳게 분류된 문서의 개수의 비율값을 의미한다. 또한 분류모델의 공정한 검증을 위해 10겹 교차검증(10-fold cross validation) 방식을 사용한다.

2. 실험 결과

표 1은 본 논문에서 제안한 하이퍼링크와 태그 기반 특징가중 법을 분류정확도 측면에서 기존 기법과 비교한 결과를 보여준다. 기존기법I은 하이퍼링크 및 태그 정보를 이용하지 않고 기준 문서에 포함된 단어만을 특징으로 취하여 분류모델을 생성한 것이며, 기존기법II는 [10]의 연구와 유사하며 기존기법I에다 태그 기반 가중치 W_T를 병합한 것이다. 제안기법I,II,III은 하이퍼링크 및 태그 정보를 융합하여 주요 특징의 빈도수를 확장한 것이다. 여기서 가중치 변환을 위해 1차, 2차, $\frac{1}{2}$차함수를 이용하며, 최종 가중치 계산을 위해 필요한 최대값 M, N은 모두 2로 설정하였다.

표에서 보는 바와 같이, 권위도 기반 가중치를 환산하

기 위해 2차함수를 적용한 경우가 평균적으로 가장 높은 분류 정확도를 보였다. 태그 정보만을 활용하여 특징가중을 수행한 기존기법II는 단순한 방법임에도 불구하고 특징가중을 수행하지 않은 기존기법I 보다 약 2% 가량의 성능 향상을 보였다. 제안기법은 기존기법II에 권위도 기반 가중을 통해 2%이상의 추가적인 성능 향상을 달성하였다. 이는 중요 태그에 출현한 특징 단어가 권위도가 높은 문서에 존재하는 경우에, 그것의 출현 빈도수를 확장 보정함으로써 분류모델의 향상에 크게 기여할 수 있음을 보여준다.

표 1. 제안기법의 분류정확도
Table 1. Classification accuracy of the proposed methods

구분	변환 범위 최대값	분류 정확도 (평균)
기존기법 I		72.81
기존기법 II	N = 2	74.90
제안기법 I (1차함수)	M = 2 N = 2	76.12
제안기법 II (2차함수)	M = 2 N = 2	77.06
제안기법 III (½차함수)	M = 2 N = 2	77.04

권위도 기반 가중치의 변환 함수에 따른 분류 정확도 측면에서는, 1차함수를 적용한 경우보다 2차 또는 ½차함수를 이용하여 변환한 결과가 1% 가량의 우세를 보였다. 이는 권위도에 따른 가중치 변환을 초기 권위도 값의 대소에 따라 동일한 비율로 변환하기보다 권위도가 낮거나 높은 문서에 포함된 특징 단어들에 대하여 상대적인 차이를 확대하여 가중치를 설정하는 것이 주효했음을 보여준다. 이러한 특성을 반영하여 2차 또는 ½차함수 이외의 함수를 적용해 보았지만 그 차이는 미미하였다.

그림 3은 권위도 가중치의 변환을 위한 2차 또는 ½차 함수에 대하여 최종 가중치 W의 최대값 N의 변화에 따른 분류정확도를 보여준다. 여기서 가로축은 최종 가중치 W의 최대값 N을 의미하고, 세로축은 분류정확도를 의미한다. 그리고 권위도 기반 가중치 W_P의 최대값 M은 2로 고정, 진출링크의 중요도인 α 값은 2차, ½차 변환 함수 각각에 대하여 -0.5, -0.1로 고정한다. 그림에서 보는 바와 같이 N값이 2.0일 때 가장 높은 정확도인 77.2%를 기록하였다. 두 함수간의 성능 차이는 크지 않았으며, 함수 특성에 따라 N값이 작을 때는 ½차함수로 변환된 경우가 우세하고, N값이 클 때는 2차함수로 변환된 경우가 우세하였다. 그리고 N값이 2.0을 초과하는 경우 정확도가 감소하는 것을 볼 수 있는데, 이는 특징가중을 통해 보정된 특징 단어들의 빈도수에 대한 차별화가 과도히 진행되어 분류모델의 성능을 저하시킨 것으로 분석한다.

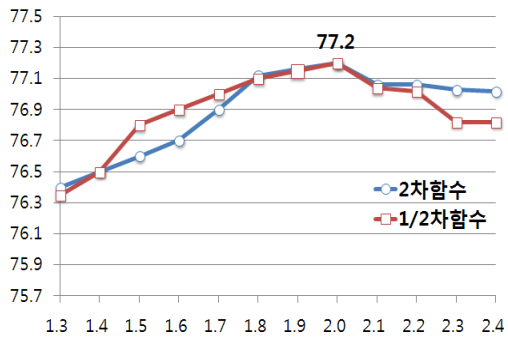

그림 3. 최종 가중치 W의 최대값 N에 따른 자동 문서 분류 정확도 (M=2 고정, $1.3 \leq N \leq 2.4$)
Fig. 3. Changes of classification accuracy from varing the maximum N of the final weight W

그림 4는 진출링크의 중요도 α 값(가로축)의 변화에 따른 자동분류의 성능 변화를 보여준다. 여기서 최종 가중치 W의 최대값 N은 그림 3의 실험 결과를 반영하여 2로 고정하였다. 2차 또는 ½차 함수를 적용하여 권위도 가중치를 변환한 경우에, 진출링크의 중요도 α가 각각 -0.5 또는 -0.1 일때 가장 높은 성능을 보였다.

그리고 α 값이 0 이상인 경우에는 분류 성능이 급격히 저하되는 것을 관찰할 수 있다. 이는 α 값이 커지면 선택되는 인접 웹문서의 수가 늘어나 결과적으로 특징 단어의 개수를 높임으로써 카테고리를 분별시키지 못하는 특징 단어가 다수 포함되기 때문이다. 본래 α 값의 범위는 -1과 0사이이지만 확실한 성능 차이를 확인하기 위해 α 값을 +0.1까지 확장하였다. 반대로 α 값이 -1로 작게 설정되는 경우에는 선택되는 인접 웹문서의 수가 감소하여 유용한 특징 단어가 제외되기 때문에 분류 성능이 저하되는 것으로 분석한다. 이와 같이 α 값에 따라 분류정확도가 유동적이므로 이를 정밀하게 자동으로 설정하는 연구가 진행될 필요가 있다.

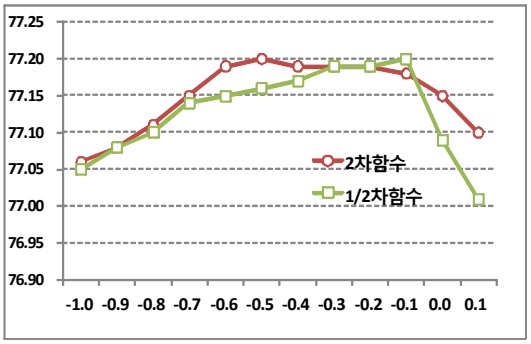

그림 4. 진출링크의 중요도 α의 변화에 따른 분류정확도
Fig. 4. Change of classification accuracy from varying the prestige of outgoing-link (α)

V. 결 론

본 논문은 웹문서의 자동분류를 위한 분류모델을 개선하기 위해 하이퍼링크 정보와 태그 정보를 융합한 특징가중 기법을 제안하였다. 최근 자동분류기술의 개선을 위해 특징 생성 및 가공 기술에 관심이 커지고 있어 웹문서에 포함된 링크 및 태그 정보를 활용하는 연구가 활발하다. 본 연구를 통해서 링크 및 태그 정보를 적극 활용하여, 특징 선택과 특징가중의 융합을 통해 초기에 산출된 특징 단어의 빈도수를 보정함으로써 자동문서분류시스템의 성능 개선에 기여했음을 확인하였다. 기존 연구의 문제점은 하이퍼링크로 연결된 인접 웹문서를 선별하지 않고 포함된 특징 단어들을 임의적으로 취한다는 것이었다. 제안 기법은 중요 태그에 포함된 특징 단어에 대하여 초기 빈도수를 보정하는 방안을 기본적으로 활용하면서, 간소화된 권위도 계산을 통해 중요하다고 판단되는 인접 문서 안에서만 특징가중을 적용하는 것이다. 결과적으로 카테고리를 차별화하는 특징 단어의 빈도수를 효과적으로 확장하여 분류모델의 성능을 높이게 되었다. 향후 연구는 HTML 태그별 가중치에 대한 최적화 기법의 적용과 진출링크 중요도 α값의 자동 설정을 위한 방안을 포함한다.

References

[1] J. Gantz, and D. Reinsel, “Extracting Value from Chaos”, http://www.emc.com/collateral/analyst-reports/, 2011

[2] Hye-young Yang, "Technology Planning Method using Big Data", Korea Institute of S&T Evaluation and Planning (KISTEP), 2012

[3] The Value and Benefits of Text Mining, JISC Digital Infrastructure, 2012

[4] J. Kim, and M. Kim, "A Study on the Implementation of SNS Message Classification by Emotion Factors", The Journal of the Institute of Internet, Broadcasting and Communiction, Vol. 11, No. 4, pp. 217-222, 2011

[5] J. Joo, and Y. Yoon, "Pattern Analysis and Prediction System for Meme Data", Journal of Korean Institute of Information Technology, Vol. 9, No. 9, pp. 163-177, 2011

[6] T.M. Mitchell, "Machine Learning", McGraw-Hill, 1997

[7] H. Altınçay, "Feature Extraction Using Single Variable Classifiers for Binary Text Classification", Lecture Notes in Computer Science, Vol. 7906, pp 332-340, 2013

[8] X. Qi, and B. D. Davison, "Web page classification: Features and algorithms", ACM Computing Surveys, Vol. 41, No. 2, Article No. 12, 2009

[9] S. Chakrabarti, B. Dom, and P. Indyk, "Enhanced hypertext categorization using hyperlinks", Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 307-318, 1998

[10] H. Utard, and J. Furnkranz, “Link-Local Features for Hypertext Classification”, Lecture Notes in Computer Science, Vol. 4289, pp. 58-69, 2005

[11] S. Brin, and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, Seventh International World-Wide Web Conference, pp. 14-18, 1998

[12] H. Benbrahim, and M. Bramer, "Impact on Performance of Hypertext Classification of Selective Rich HTML Capture", Artificial Intelligence Applications and Innovations (AIAI-2004), pp. 22-27, 2004

※ 이 논문은 2012년도 서울시립대학교 교내학술연구비에 의하여 연구되었음.

저자 소개

이 아 람(정회원)

- 2012년 : 삼육대학교 자연과학대학 컴퓨터학부 졸업 (이학사)
- 2013년 : 서울시립대학교 전자전기컴퓨터공학부 대학원 졸업 (공학석사)

<주관심분야> 데이터마이닝, 기계학습, 자연어처리

현 만(준회원)

- 2011년 : Department of Computer Science and Technology, Shenyang Ligong University, China 졸업 (이학사)
- 2012년~현재 : 서울시립대학교 전자전기컴퓨터공학부 석사과정

<주관심분야> 데이터마이닝, 기계학습, 빅데이터분석

김 한 준(정회원)

- 1994년 : 서울대학교 계산통계학과 졸업 (이학사)
- 1996년 : 서울대학교 전산과학과 대학원 졸업 (이학석사)
- 2002년 : 서울대학교 컴퓨터공학부 대학원 졸업 (공학박사)
- 2002년~2002년 : 서울대학교 공과대학 Post-Doc
- 2002년~현재 : 서울시립대학교 전자전기컴퓨터공학부 부교

<주관심분야> 데이터마이닝, 정보검색, 기계학습, 데이터베이스